人工智能场景下
语言偏见可视化工具 DADD 对不平等现象的度量

李俊麒
（上海交通大学，上海 200240）

摘　要：人工智能的发展给智能生活带来便利的同时，用户可能会因算法自动处理的个人数据而受到不公平地对待，由此产生了数字歧视这一新型不平等现象。数字歧视是算法通过继承先前决策者的偏见或复制现实世界中的歧视实例进行计算，可能导致之前处于弱势的群体受到更不公正的待遇，从而加剧现有的不平等。本文以红色药丸论坛为例，运用语言偏见可视化工具 DADD 分析其性别不平等程度，并就数字歧视这一问题提出思考。

关键词：人工智能；数字歧视；DADD；红色药丸论坛；可视化工具　　**中图分类号：**G220　　**文献标识码：**A
文章编号：1671-0134（2022）03-027-03　　DOI：10.19483/j.cnki.11-4653/n.2022.03.007
本文著录格式：李俊麒 . 人工智能场景下语言偏见可视化工具对不平等现象的度量 [J]. 中国传媒科技，2022（03）：27-29.

导语

2019 年，联合国教科文组织发布了一篇报告名为《I' d blush if I could》，揭示了人工智能研发和应用中的性别差距和性别偏见问题。[1]AI 语音助手包括亚马逊的语音助手 Alexa 和苹果语音助手 Siri 等几乎所有语音助手都是女性的名字。[2] 阿里、小米、百度 AI 语音助手默认语音都是温柔悦耳的女声，她们几乎都被设定为相同的、谦逊和顺从的女性形象。2014 年，亚马逊将过去 10 年收到的简历作为数据库，开发了一款筛选简历的 AI 工具，该工具将含有“女性”的所有简历降级。2019 年 11 月，苹果和高盛共同打造的 Apple Card 因在设定配额算法时涉嫌性别歧视而被美国监管机构调查。

对于出现的这个现象，斯坦福大学教授 Brian Arthur 在《技术的本质》一书中提到，“无论我们是否注意到它，在我们历史的这个阶段，科技已经让人们感到压抑和困扰，算法技术和人工智能的高速发展正在威胁着性别平等的共识愿景。”[3] 算法日益成为信息传播的主力军，然而其性别歧视潜力变得越来越明显。算法开发的简化特性忽视了女性社会的多样性，难以避免地产生了技术先存偏见和数据偏见。传统媒体对女性的“男性凝视”（Male gaze）演变为算法传播的“代码凝视”（Code gaze），数字歧视日益严重。由此可见，算法性别歧视的治理作为新时代的科技伦理问题亟待关注。

1. 数字歧视

数字歧视（Digital discrimination）是指基于算法自动决策形成的直接或间接的歧视行为。日常生活中，越来越多的决策被委托给算法，从申请的工作到购买的产品、阅读的新闻以及浏览的网页，越来越多的重要决定默认委托给算法系统进行处理。有时候算法做出的自动化决策，包括基于机器学习的决策，被认为是完美无缺的，没有人类的大部分缺点（例如疲劳或个人偏见）。并且与人类做出的决定相比，算法做出的决定较少经过仔细审查。然而，自动化决策，尤其是机器学习算法，很可能会继承程序员以前的决策偏见、用户偏见或社会偏见，这会导致歧视性结果。

目前国外学者对数字歧视的研究和证明主要集中在性别歧视、种族歧视、收入歧视、地域歧视等方面，而国内在这个领域的研究还比较欠缺，因此文章将以性别歧视为例，以红色药丸论坛为案例研究对象，运用语言偏见可视化工具探究词汇嵌入模型能在多大程度上追踪性别偏见，以揭露该论坛中的数字性别歧视程度，并对研究结果进行分析和批判性思考。

2. 红色药丸论坛

红色药丸是 Reddit 上的一个在线论坛，于 2012 年 10 月建立，超过 30 万人订阅了该频道。该论坛的名称来自一部名 The Matrix 的电影。故事的主角被要求在蓝色药丸和红色药丸之间做出选择。如果他选择蓝丸，他将继续舒适但虚假的生活；如果他选择红色药丸，他将身处真实的但更黑暗的世界。最终，他吞下红色药丸，并承认自己生活在一个有许多真相并奴役着他的世界。红色药丸论坛旨在揭示女权主义的“真实本质”，即女权主义只是压制男性的压迫工具。它的订阅者认为该论坛是在男性缺乏积极认同日益严重的文化中讨论性别策略的场所，在一定程度上帮助男性在社会中重新获得应有的地位。[4]

论坛主要采用的是 KARMA 算法，用户可以在红色药丸论坛中创建话题，且针对感兴趣的各种主题发起话题讨论。用户通过发表评论并在一个帖子上投赞成票或反对票来增加或减少该帖子的分数，具有更高分数的帖子更容易被其他用户看到，而算法极少推送分数低的帖子。投票系统引导着热门用户主导对话，同时防止各种对论坛的负面想法。此外，话题的版主为论坛创建了行为准则，对用户

在该话题下发表的观点进行奖励或制裁。对红色药丸论坛主流价值观持积极态度的用户将成为“红色药丸精英”，版主通过赋予其特殊标志向其他用户展示该用户的态度和立场；而对红色药丸论坛主流价值观持反对意见的用户将被阻止再次访问论坛。因此，版主会尽可能选择最适合讨论红色药丸主题的合格用户，通过给予奖励来激励他们不断输出符合红色药丸主题的观点。

男性权利是红色药丸论坛最受欢迎的话题。该话题鼓吹男性正面临着诸多不平等的状况。性别意识形态在于群体中的个体试图获得该群体其他成员的社会认可。如果这些个体更极端，那么他们将获得更多来自这个群体的认可，导致群体接受越来越多的极端意见，从而激化性别偏见。近年来，红丸论坛中存在大量对女性的极端看法，已成为滋生性别对立的温床，因此论坛中可能存在着数字歧视。本文通过语言偏见可视化工具分析该论坛是否存在基于性别的数字歧视，以及歧视程度如何。

3. 语言偏见可视化工具

3.1 概述

人工智能的发展让机器深度学习成为发现语言偏见的重要手段，该领域最著名的工具之一就是 DADD 语言偏见可视化工具（Language Bias Visualiser Discovering and Attesting Digital Discrimination）。DADD 通过文本嵌入模型（Words Embedding Model）从互联网上捕获数据集，以交互的方式处理庞大文本数据集中的男性和女性固有偏见，追踪和汇总不同形式的数字歧视，从数据库中的用户数据中检测含有潜在或隐含偏见的文本，将文本转换为高维向量，捕捉文本之间的语义关系，然后通过聚类算法对数据集进行分类，再通过语义分析系统标记更多的概念偏见。[5] 本文对红色药丸的文本数据集中的偏见进行了偏见词频率、词语的偏见程度、偏见词的分布和文本偏见极性分析，深入了解该群体中存在的偏见。

3.2 偏见词分析

3.2.1 最常见的偏见词

下面的词云图片展示了红色药丸论坛中对男性和女性最常用的偏见词汇。单词越黑越大，说明该单词出现的频率越高。



3.2.2 偏见程度最高的词

下表对比了从 300 个最具有偏见的词中选出的前 10 个男性偏见词和女性偏见词。Bias 表示对每个性别的偏见程度，范围从 1（强偏见）到 0（无偏见）。由下表可知对女性偏见度的平均值为 0.849，高于男性的偏见度的 0.764，这说明对女性的偏见程度要明显大于男性。

Male		Female	
Word	Bias	Word	Bias
himself	0.9516	chick	0.9909
businessman	0.8046	hb	0.9388
hero	0.7785	plate	0.8939
leader	0.7716	fwb	0.8586
warrior	0.7488	ons	0.822
tyler	0.7485	flakes	0.8127
badass	0.7435	dtf	0.8122
donald	0.7139	slutty	0.7925
putin	0.6914	gal	0.7865
hitler	0.6902	chicks	0.7804

表 1 偏见程度最高的词对比表（Top10）

3.2.3 文本嵌入模型

下面的散点图显示了单词的嵌入表示。它们是由文本嵌入模型学习，训练红色药丸论坛中的所有文本以获得结果，同时使用 t-SNE 将所有嵌入转换为两个主要维度。语义相关的词在图中挨得很近，语义无关的词相距更远。每个单词的大小与其在数据集中的频率有关，出现频率更高的词周围的圆圈更大，显示出了嵌入空间中男性和女性偏向词的分布。对男性来说，偏见词更集中在第一和第四象限，相反对女性来说，在第二和第三象限中可以找到更多的偏见词。

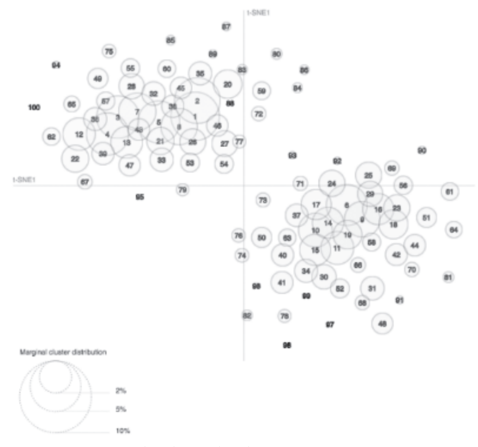


图 2 男（绿）女（橙）嵌入空间图对比图

3.2.4 文本偏见极性

文本见偏极性（Words Bias Polarity）表示文字的情感色彩，分为 7 个等级。数字从 0.6 到 1 表示非常正面；0.3 到 0.6 表示正面；0.005 到 0.3 表示轻微正面；–0.005 到 0.005 表示中立；–0.3 到 –0.005 表示轻微负面；–0.6 到 –0.3 表示负面；–1 到 –0.6 表示非常负面。下面两个饼图显示了从红色药丸论坛中 300 个最有性别偏见的词的语言偏见极性占比。

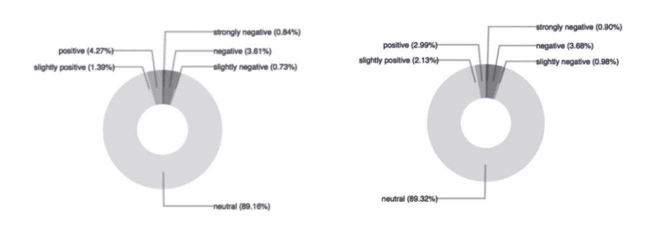


图 3　男（左）女（右）语言偏见极性占比

对于男性，89.16% 的词是中性的。非常正面、正面和轻微正面的词分别占 0.84%、4.27% 和 1.39%，而轻微负面、负面和强烈负面的词分别为 0.73%、3.61% 和 0.84%；对于女性，80.32% 的词是中性的，非常正面、正面词和轻微正面的词分别占 0.9%、2.99% 和 2.13%，而轻微负面、负面和强烈负面分别为 0.98%、3.68% 和 0.90%。由图可见，男性的正面词总占比要高于女性，女性受到更多负面偏见的影响。因此，红色药丸论坛中存在数字歧视，女性更容易受到负面偏见。

4. 总结与讨论

通过语言偏见可视化工具 DADD，本研究发现了红色药丸论坛中存在的偏见概念，找出了最具象征意义的概念词和学习了有偏见产生的过程。通过查找、追踪和分析歧视词，将其可视化之后从而准确而清晰地对社会问题有更深入地理解和判断。从横向上看，本研究量化了性别偏见词的数量和范围，从而可以比较红色药丸论坛和其他论坛之间的歧视水平；从纵向来看，在不同的角度测试红色药丸论坛的性别偏见时会得到不同的结果，体现出性别偏见的发展趋势如何。雨果曾说“俚语是语言中最多变但最重要的部分”。偏见词可以反映一些社会问题，尤其是这些词背后存在的各种歧视。本研究仅以性别歧视为例进行了分析，但社会上还存在其他如种族、收入、地域歧视等，这些歧视通常以不同的词语呈现出来，这也将成为未来继续探索的方向。

根据本研究，红色药丸论坛中男性存在着女性的偏见，背后的深层次原因值得反思。为什么网络平台的性别偏见如此之大？它的核心逻辑是什么？学者 Amelita 认为大多数用户只是孤独、年轻或脆弱。[6] 严格的监管对局外人十分敌视，因此红色药丸论坛上的有毒信仰更容易引起一见钟情的反感和仇恨。又因为他们太年轻，还没有形成自己的价值观，很容易受极端言辞的影响。因此，相关主体该采取一些必要措施来减少用户的数字性别歧视行为。

一是管理部门的监管。相关管理部门应该对网络社区进行合理的监管，及时关闭宣扬极端思想的社群，惩罚表达极端思想的用户。同时还要加强性别教育，引导公民对性别有正确的认识。[7] 二是算法设计的优化。一方面在算法设计过程中，算法工程师应该全方位收集数据，增加数据的完整性，使其能够充分体现男性和女性的社会生存状况，避免大数据过度呈现造成的偏见。另一方面，可以在编码过程中设计更具包容性的代码，通过及时发现用户在互动中呈现的歧视行为，及时纠正已出现的歧视形式。[8] 三是人工智能技术的透明化。通过明确用户的算法机制有助于社会监督确认算法是否带有偏见，以消除受众对算法技术的过度信任或不信任；同时也有助于明晰责任范围，追溯责任主体，提高算法工程师和相关平台的责任感。[9] 媒

参考文献

[1] 联合国教科文组织 .I’ d blush if I could：closing gender divides in digital skills through education[EB/OL].[online] Available at：<https：//unesdoc.unesco.org/ å >,2019.

[2] 汪怀君 . 人工智能消费场景中的女性性别歧视 [J]. 自然辩证法通讯 ,2020（5）:45-51.

[3] Brian，A.The Nature of Technology：What It Is and How It Evolves[M].2011.Free Press.

[4] Pierce D，Misogynistic Men Online： How the Red Pill Helped Elect Trump Journal of Women in Culture and Society.vol.44，no.3.2019.

[5] Ferrer，X.and Van Nuenen，T.，n.d.Language Bias Visualiser-DADD.[EB/OL].https：//xfold.github.io/WE-GenderBiasVisualisationWeb/.

[6] Amelita T.https：//www.newstatesman.com/science tech/internet/2017/02/reddit-the-red-pill-interview-how-misogyny-spreads-online.2011.

[7] 宋素红，王跃祺，常何秋子 . 算法性别歧视的形成逻辑及多元化治理 [J]. 当代传播 ,2020（5）:95-100.

[8] Wihbey，J.The possibilities of digital discrimination：Research on e-commerce，algorithms and big data. Journalists resource，2015.

[9] 昌沁 . 新闻传播中人工智能技术造成的伦理失范与对策 [J]. 中国传媒科技，2020（11）：28-30.

作者简介：李俊麒（1998-），女，四川宜宾，上海交通大学文创学院、伦敦大学国王学院数字人文系双硕士，研究方向：文化创意产业、数字人文、文化与社会大数据。

（责任编辑：张晓婧）